A Statistical Approach for Recognizing Emotion from Dance Sequence

HanHoon Park[1], Jong-Il Park[1], Un-Mi Kim[2], Woontack Woo[3]
[1] Department of Electrical and Computer Engineering, [2] Department of Dance,
Hanyang University,
17 Haengdang, Seongdong, Seoul, 133-070, Korea
Tel. +82-2-2299-7820, Fax.: +82-2-2299-7820
[3] U-VR Lab., K-JIST,
Oryong-dong, Puk-gu, Kwangju, 500-712, Korea
Tel. +82-62-970-3157, Fax.: +82-62-970-3245
e-mail : [hanuni, jipark]@mr.hanyang.ac.kr, wwoo@kjist.ac.kr

Abstract: We propose a simple method that can recognize human emotion from monocular dance image sequences. The method only exploits the information within image sequences and does not require cumbersome attachments like sensors. This makes the method a simple, human-friendly one. Moreover, the method is more robust and efficient by taking into account the statistical property of image sequences based on PCA (Principal Component Analysis). The correct recognition rate in real-time is about 75% in a variety of experiments.

1. Introduction

The last couple of years have seen that many researchers focused on recognizing human emotion to achieve a more efficient and natural way of human-computer interface.

Picard tries to recognize affective state by capturing the physiological signals, e.g. EMG (electromygram), blood volume pressure, galvanic skin response, and respiration etc [1]. Picard considered the emotion recognition as physiological pattern recognition. However, it needs complicated devices to capture the physiological signals.

Toward more human-friendly interface, the emotion recognition methods based on computer vision techniques have been proposed [2,3]. These methods extract emotional information from speeches or facial expressions. However, these methods didn't focus on gestures to recognize human emotion because the gestures are too much high dimensional and dynamic, and moreover it's not easy to track them exactly. Recently, some methods that can track the gestures exactly were introduced [4,5], but they are too complex to be applied to real-time systems for recognizing human emotion.

Figure 1. Rectangle surrounding human body.

Woo et al. proposed the method that can recognize emotion from dance image sequences simpler and faster [6,7]. They simplified the high-dimensional and dynamic change of gestures to the movement of rectangle surrounding human body (Figure 1). Based on Laban's theory [8], they extracted the features; the width and height of rectangle, the coordinate of centroid, etc., and classified them into predefined emotional category using TDMLP.

The method in this paper is an extension of the previous work proposed by Woo et al. We attempt to recognize human emotion faster and more correctly from dance image sequences.

In general, the features extracted from the movement of rectangle are still high dimensional and dynamic. Thus, directly extracting emotional information from all the features is difficult and sensitive to noise. The proposed method tries to be robust to noise while maintaining high recognition rate by classifying the extracted features in subspace using PCA.

The rest of this paper is organized as follows. In the next section we explain the techniques for recognizing human emotion from dance image sequences. In section 3 we mention several issues raised in implementing the proposed method and show our solutions for them. In section 4 we provide the experimental result. We present conclusions in section 5.

2. Emotion Recognition from Dance Image Sequences

The proposed method consists of three parts:

- Feature extraction from dance image sequence;
- Extraction of principal components from the features using PCA;
- Classification of the principal components into predefined emotional space using neural network.

2.1 Feature extraction from dance image sequence

At first, we get rid of the background and shadow of dance images using difference keying and normalized difference keying technique separately and create binary images (Figure 1)[9]. Next, we extract the features representing dancing motion, e.g. the width and height of rectangle, the coordinate of centroid, silhouette area, the ratio between rectangle and silhouette area, and the velocity and acceleration of each feature (Table 1).

Table 1. Features extracted from binary images

The ratio between rectangle and silhouette area	W
The coordinate of centroid	(C_x, C_y)
The coordinate of the center of rectangle	(R_x, R_y)
The silhouette area	S_s
The rectangle area	S_r
The velocity of each feature	f(·)
The acceleration of each feature	g(·)
$f(x_n) = x_n - x_{n-1}$, $g(x_n) = x_n - 2*x_{n-1} + x_{n-2}$	

2.2 Extraction of principal components

We applied SVD (Singular Value Decomposition) to the extracted features and selected the principal components having large eigen value. This has an effect to remove noisy information from the features [9].

In case of extracting n features from the dance sequence having m frames, we apply SVD to **A** matrix having m×n features as its elements. That is:

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T$$

where

$$\Sigma = \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \quad \mathbf{D} = \begin{pmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r \end{pmatrix}$$

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0, \qquad r \leq m, n.$$

Here, σ_i means the singular value and ith column of U is eigen vector concerned with σ_i. The r eigen vectors ($\mathbf{u}_1$, $\mathbf{u}_2$, …, $\mathbf{u}_r$) concerned with large σ_i are represented by linear combination of the feature vectors ($\mathbf{f}_1$, $\mathbf{f}_2$, …, $\mathbf{f}_n$). That is:

$$\mathbf{u}_1 = \alpha_{11}\mathbf{f}_1 + \alpha_{12}\mathbf{f}_2 + \cdots + \alpha_{1n}\mathbf{f}_n ,$$
$$\mathbf{u}_2 = \alpha_{21}\mathbf{f}_1 + \alpha_{22}\mathbf{f}_2 + \cdots + \alpha_{2n}\mathbf{f}_n ,$$
$$\vdots \qquad\qquad \vdots$$
$$\mathbf{u}_r = \alpha_{r1}\mathbf{f}_1 + \alpha_{r2}\mathbf{f}_2 + \cdots + \alpha_{rn}\mathbf{f}_n .$$

This is rewritten as follows:

$$\mathbf{u}_i = \mathbf{F}\alpha_i \quad for\ i = 1, 2, \cdots, r,$$

where

$$\alpha_i = \begin{bmatrix} \alpha_{i1} & \alpha_{i2} & \cdots & \alpha_{in} \end{bmatrix}^T ,$$
$$F = \begin{bmatrix} f_1 & f_2 & \cdots & f_n \end{bmatrix}$$

Here, α_i represents contribution measure and it's solved as follows:

$$\alpha_i = (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{u}_i .$$

Given α_i the principal components are represented by weighted sum of the features from dance image sequence:

$$p_1 = \alpha_{11}f_1' + \alpha_{12}f_2' + \cdots + \alpha_{1n}f_n' ,$$
$$p_2 = \alpha_{21}f_1' + \alpha_{22}f_2' + \cdots + \alpha_{2n}f_n' ,$$
$$\vdots \qquad\qquad \vdots$$
$$p_r = \alpha_{r1}f_1' + \alpha_{r2}f_2' + \cdots + \alpha_{rn}f_n'$$

where f_i' represents the features extracted from each frame of dance sequence.

2.3 Classification of the principal components

As a kind of neural network, MLP (Multi-Layer Perceptron) is feedforward network that have hidden layers between input layer and output layer. MLP can be used to classify arbitrary data because its internal neurons have nonlinear property. But it can't classify dynamically changing data like dance. As shown in Figure 2, TDMLP (Time Delay MLP) stores the input data for some period and uses all the delayed data as input. Because it can effectively analyze not only an instant value but also changing patterns of input data, it can classify the dynamic data exactly and robustly.

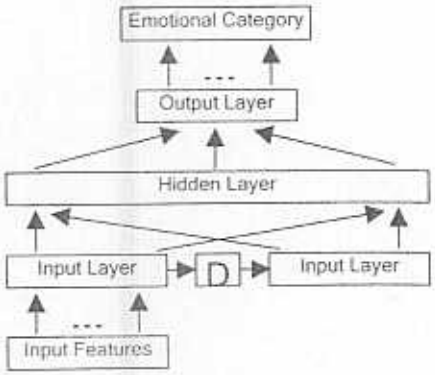

Figure 2. The structure of TDMLP. It has three layers and a buffer in input layer.

3. Implementation Issues and Solutions

In general, the features extracted from dynamically changing dance images include many outliers. Applying SVD to the extracted features without getting rid of these outliers, the recognition shows poor results due to the outliers. To resolve this problem, we calculate the mean and deviation of the features in advance and clamp the outliers that deviate from the mean more than an allowed deviation. In addition, the features are too complexly mixed each other to be classified by TDMLP. To increase the separability of the features, we normalize them to be distributed evenly between the maximum value and minimum value of the features.

```
//Data clamping
if( (data – average) > (max - average)*α)
      data = average + (max - average)*α;
else if( (data – average) < (min - average)*β)
      data = average + (min - average)* β;

//Data normalization
data = (data - min) / (max - min);

max : maximum value of data
min: minimum value of data
average : average value of data
α, β : coefficients determined according to the complexity of data
```

Figure 3. The source code for clamping and normalizing data.

The features are very noisy and need to be smoothed to obtain a reasonable result. Thus, we take median filtering. The window size for median filtering heavily affects recognition rate. If the window size is too small, it tends to be sensitive to noise. On the contrary, if the window size is too large, the separability of the features is deteriorated. Figure 4 shows how the size of the window for median filtering affects the recognition rate.

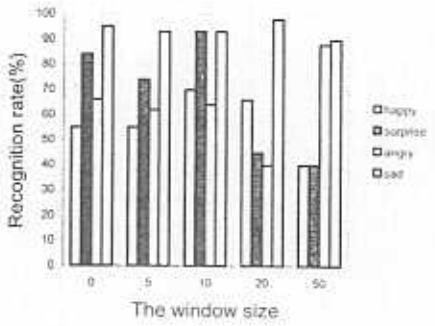

Figure 4. Recognition rate according to the window size for median filtering.

Because the SVD analyzes the statistical property of the features, the training sequences should contain sufficient number of patterns. Insufficient training data may result in erroneous recognition. To resolve this problem, we captured the dance image sequence including various dancing motion for a long time.

The physical difference of dancer can cause a problem. Despite expressing the same motion, the features extracted from one dancer to another may be different and thus it can be recognized as different one. To resolve this problem, we individually normalize the extracted features for each dancer.

4. Experiments and Results

To obtain experimental sequences, we captured the dance motion from 4 professional dancers using a video camera (Cannon MV1). Two sequences are used for training the TDMLP and the others are used for testing the TDMLP.

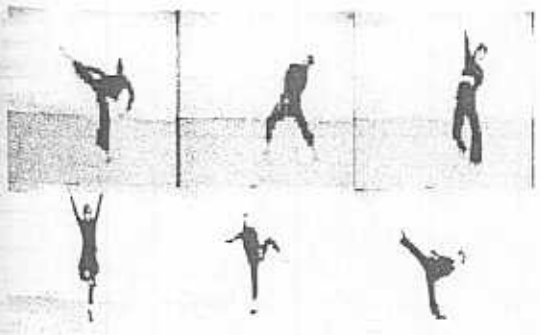

Figure 5. Dance images.

At first, we eliminate the background and shadow of each frame in sequences and get binary images. Next, we extract 21 features representing dancing motion and apply SVD to them. We calculate the contribution measure for 7 eigen vectors having large eigen values and extract 7 principal components every frame. Finally, we buffer 14 principal components with one delay and exploit them as the input of the TDMLP. The TDMLP is learned to map the principal components to predefined emotional category (happy, sad, angry, surprise) and is tested to recognize the emotion that an arbitrary image sequences represent. The TDMLP consists of 14 input nodes, 56 hidden nodes, and 4 output nodes.

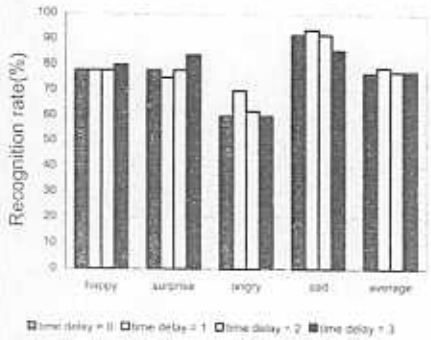

Figure 6. Recognition rate according to the number of time delays.

Figure 6 shows the recognition rate of 4 emotions to the number of time delay of TDMLP. Our system shows 75% recognition rate on average. It shows a little difference according to the number of time delay of TDMLP. As shown in Figure 6, our system shows best performance when time delay = 1.

Table 2. Recognition rate for each emotion (inside the training sequence)

	happy	surprise	angry	sad
happy	79%	11%	8%	2%
surprise	3%	78%	11%	8%
angry	23%	0%	70%	7%
sad	0%	7%	1%	92%

Table 3. Recognition rate for each emotion (outside the training sequence)

	happy	surprise	angry	sad
happy	52%	16%	24%	8%
surprise	16%	76%	0%	8%
angry	14%	12%	64%	10%
sad	0%	4%	2%	94%

Table 2 and 3, when time delay = 1, show the recognition rate for each emotion to inside and outside the training sequence separately. The recognition rate in outside the training sequence is less than that in inside the

training sequence since the method of each professional dancer's expressing his/her emotion is different. It can be alleviated if we acquire the sample sequence from more dancers and have the TDMLP learn to adapt to various expression methods included.

Table 4. Recognition rate according to the number of frames in training sequences.

frames	happy	surprise	angry	sad
1000	30%	0%	0%	100%
2000	30%	55%	25%	100%
3000	50%	76%	60%	70%
4000	50%	50%	64%	85%
5000	52%	76%	64%	94%

Table 4 shows how the number of frames in the training sequences affects the recognition rate when time delay = 1 (outside the training sequence). If the sequence is too short, it has a few specific patterns. The TDMLP learned with the sequence can recognize only the specific patterns and can't have good recognition rate. When frames > 5000, our system had shown satisfactory performance.

5. Conclusion

We proposed a real-time emotion recognition method taking into account statistical properties of the features extracted from dance image sequences. We verified through experiments that it shows reasonable performance for practical application.

The proposed method can be practically used in entertainment and education fields that need humanlike, real-time human-computer interaction on which we are intensively investigating.

The important avenue for future work will be to capture the motion from more dancers and recognize their emotions.

References

[1] R. Picard, E. Vyzas, J. Healey, "Toward machine emotional intelligence: analysis of affective physiological state," *IEEE Trans. PAMI*, Vol. 23, No. 10, pp. 1175-1191, Oct. 2001.

[2] R. Nakatsu, J. Nicholson, N. Tosa, "Emotion recognition and its application to computer agents with spontaneous interactive capabilities," In *Proc. ICMC*, vol 2, pp. 804-808, 1999.

[3] L. S. Chen, T. S. Huang, "Emotional expressions in audiovisual human computer interaction," In *Proc. of ICME*, vol 1, pp. 423-426, 2000.

[4] M. Isard, A. Blake, "CONDENSATION – conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29, no. 1, pp. 5-28, 1998.

[5] T. J. Cham, J. M. Rehg, "A multiple hypothesis approach to figure tracking," In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 239-245, Fort Collins, Colorado, 1999.

[6] R. Suzuki, Y. Iwadate, M. Inoue, W. Woo, "MIDAS: MIC Interactive Dance System", *IEEE Int'l conf. on Systems, Man and Cybernetics*, vol. 2, pp. 751-756, Oct. 2000.

[7] W. Woo, J. Park, Y. Iwadate, "Emotion analysis from dance performance using time-delay neural networks," In *Proc. of CVPRIP*, vol. 2, pp. 374-377, Feb. 2000.

[8] R. Raban, *Modern educational dance*, Trans-Atlantic Publications, Inc. 1988.

[9] H. Park, J. Park, W. Woo, "Realtime emotion recognition from dance image sequence using TDMLP," In *Proc. HCI 2002*, Feb. 2002.